When is Constrained Clustering Beneficial, and Why? [*]

Kiri L. Wagstaff
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena CA 91109
kiri.wagstaff@jpl.nasa.gov

Sugato Basu
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
basu@ai.sri.com

Ian Davidson
Department of Computer Science
State University of New York, Albany
1400 Washington Ave., Albany, NY 12222
davidson@cs.albany.edu

Abstract

Several researchers have shown that constraints can improve the results of a variety of clustering algorithms. However, there can be a large variation in this improvement, even for a fixed number of constraints for a given data set. We present the first attempt to provide insight into this phenomenon by characterizing two constraint set properties: informativeness and coherence. We show that these measures can help explain why some constraint sets are more beneficial to clustering algorithms than others. Since they can be computed prior to clustering, these measures can aid in deciding which constraints to use in practice.

Introduction and Motivation

The last five years have seen extensive work on incorporating instance-level constraints into clustering methods (Wagstaff *et al.* 2001; Klein, Kamvar, & Manning 2002; Xing *et al.* 2003; Bilenko, Basu, & Mooney 2004; Bar-Hillel *et al.* 2005). Instance-level constraints specify that two items must be placed into the same cluster (must-link, ML) or different clusters (cannot-link, CL). This semi-supervised approach has led to improved performance on several real-world applications, such as noun phrase coreference resolution and GPS-based map refinement (Wagstaff *et al.* 2001), person identification from surveillance camera clips (Bar-Hillel *et al.* 2005) and landscape detection from hyperspectral data (Lu & Leen 2005).

However, the common practice of presenting results using learning curves, which average performance over multiple constraint sets of the same size, obscures important details. For example, we took four UCI data sets (Blake & Merz 1998) and generated 1000 randomly selected constraint sets, each containing 25 constraints. We then clustered each data set with COP-KMeans (Wagstaff *et al.* 2001), using the same initialization for each clustering run. We observed that, even when the number of constraints is fixed, the accuracy of the output partition measured using the Rand Index (Rand 1971) varies greatly, by 6 to 10% (Table 1). Since the starting point for each run was held fixed, the only source of variation was the constraint set itself.

Data set	**Constrained Accuracy** Min	Mean	Max	**Unconstrained** Mean
Glass	66.3	70.3	74.0	69.0
Ionosphere	56.6	59.3	62.6	58.6
Iris	83.6	88.2	93.4	84.7
Wine	67.4	70.8	74.5	70.2

Table 1: Variation in accuracy (Rand Index) obtained by COP-KMeans using 25 randomly generated constraints and a fixed starting point, over 1000 trials.

In fact, we found that although the average constrained accuracy exceeds that of the average unconstrained accuracy, as expected, we observe that the constrained results can produce results that are significantly worse than not using constraints at all (compare “min” column to rightmost column in Table 1). This occurs even though the constraints are noise-free. Our focus in this work has been to explain this phenomenon and to offer ways to estimate the utility of a given constraint set. We have identified two constraint set properties that help explain these variations: *informativeness* and *coherence*.

Quantifying Informativeness and Coherence

Informativeness refers to the amount of information in the constraint set that the algorithm cannot determine on its own. Given an algorithm $\mathcal{A}$, we generate partition $P_{\mathcal{A}}$ by running $\mathcal{A}$ on the data set without any constraints. We then calculate the fraction of constraints in constraint set $\mathcal{C}$ that are unsatisfied by $P_{\mathcal{A}}$. If every constraint in $\mathcal{C}$ can be satisfied by $\mathcal{A}$’s default behavior, then $\mathcal{C}$ has 0 informativeness for $\mathcal{A}$. On the other hand, if $\mathcal{C}$ contains several constraints that $\mathcal{A}$ cannot guess on its own, then it is very informative.

Coherence is the amount of agreement between the constraints in set $\mathcal{C}$, given a distance metric $\mathcal{D}$. It is not algorithm dependent. An ML (or CL) constraint can be viewed as imposing an attractive (or repulsive) force in the feature space within its vicinity. Two constraints are perfectly coherent if they are orthogonal to each other and incoherent if they are parallel to each other. To determine the coherence of two constraints, c_1 and c_2, we compute their *projected overlap*, or how much the projection of c_1 along the direction of c_2 overlaps with (interferes with) c_2. We define coherence as the fraction of ML–CL constraint pairs in the set that have zero projected overlap.

[*] This work was supported by NSF ITR #0325329 and DARPA (CALO Project: Contract #NBCHD030010, Order #T310), and it was performed in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

	Mean performance gain/loss				Informativeness				
Data Set	CKM	PKM	MKM	MPKM	$\mathcal{I}_{CKM}$	$\mathcal{I}_{PKM}$	$\mathcal{I}_{MKM}$	$\mathcal{I}_{MPKM}$	$\mathcal{COH}$
Glass	1.3	25.2	17.0	28.3	0.28	0.44	0.51	0.50	0.70
Ionosphere	0.7	-1.0	0.0	-1.0	0.41	0.41	0.42	0.42	0.65
Iris	3.5	4.0	5.5	2.9	0.12	0.12	0.11	0.11	0.94
Wine	0.6	0.4	-1.9	-2.6	0.30	0.27	0.06	0.06	0.77

Table 2: Average performance gain (or loss) for four constrained clustering algorithms, using 1000 randomly generated 25-constraint sets. The right side of the table reports the average informativeness ($\mathcal{I}$) and coherence ($\mathcal{COH}$) values of these sets.

Experimental Results

To understand how these constraint set properties affect various algorithms, we conducted the same experiment with 1000 randomly generated 25-constraint sets using four different constrained clustering algorithms. The two most common approaches to constrained clustering involve either satisfying the constraints directly or learning a distance metric that accommodates the constraints. We compared a representative of each approach and a hybrid method that performs both functions: (1) COP-KMeans (CKM): hard constraint satisfaction in KMeans (Wagstaff *et al.* 2001); (2) PC-KMeans (PKM): soft constraint satisfaction (Bilenko, Basu, & Mooney 2004); (3) M-KMeans (MKM): metric learning from constraints (Bilenko, Basu, & Mooney 2004); and (4) MPC-KMeans (MPKM): hybrid approach, performing both soft constraint satisfaction and metric learning (Bilenko, Basu, & Mooney 2004).

First, we report the mean performance gain (or loss) in terms of Rand Index that was achieved by each algorithm with each data set (left side of Table 2). We see that even the average results can indicate a negative impact from the constraints, for some algorithms. PKM, MKM, and MPKM attain very large improvements when using constraints with the Glass data set (17–28%) because their default unconstrained performance is very low. The right side of Table 2 shows the average informativeness and coherence for each algorithm and data set.[1] We see that the large increases in performance for PKM, MKM, and MPKM for the Glass data set correspond to high informativeness values (higher than that for CKM with Glass, which demonstrates only a modest increase in accuracy). However, high informativeness is not sufficient for predicting accuracy improvement, as the results for Ionosphere indicate. The Ionosphere constraints, although informative, also tend to have lower coherence than any other data set's constraints. Incoherent sets are difficult to completely satisfy, and we see this reflected in the lack of significant improvement when using constraints with this data set. Conversely, the Iris constraints have very high coherence (0.94) but low informativeness, leading to the modest (but positive) average effect on performance for all algorithms. The Wine constraints have a remarkable lack of informativeness for MKM and MPKM, so even though the coherence of the constraint set is reasonably high (0.77), the 23% of constraint pairs that are incoherent dominates performance and explains the small decrease in average accuracy.

Conclusions and Future Work

We have proposed two measures of constraint set properties, informativeness and coherence, than can help explain the variations in constrained clustering behavior we observe. Identifying meaningful constraint set properties is of benefit to practitioners and researchers. For scenarios in which the user can generate multiple constraint sets, these results recommend selecting the set with the highest informativeness and coherence values to avoid situations in which the constraints may negatively impact performance. Our measures can also potentially be used to prune noisy constraints or to actively choose constraints. We intend to explore these options in the future. We also plan to generalize our definition of coherence to non-metric distance measures and to explore the use of coherence as a guide in selecting the most appropriate distance metric for a given data set. Further, these measures can provide insight into the black-box computation of different metric-learning constrained clustering algorithms. Since these methods modify the distance metric, the concept of coherence helps explain why they are so effective: they implicitly increase coherence as they iterate. This effect is now quantifiable, and an experimental evaluation of how coherence changes as the algorithm iterates is likely to increase our understanding of metric-learning methods.

References

Bar-Hillel, A.; Hertz, T.; Shental, N.; and Weinshall, D. 2005. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research* 6:937–965.

Bilenko, M.; Basu, S.; and Mooney, R. J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *21st Int'l Conf. on Machine Learning*, 11–18.

Blake, C. L., and Merz, C. J. 1998. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Klein, D.; Kamvar, S. D.; and Manning, C. D. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *19th Int'l Conf. on Machine Learning*, 307–313.

Lu, Z., and Leen, T. K. 2005. Semi-supervised learning with penalized probabilistic clustering. In *NIPS 17*.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *JASA* 66(366):846–850.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained k-means clustering with background knowledge. In *18th Int'l Conf. on Machine Learning*, 577–584.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2003. Distance metric learning, with application to clustering with side-information. In *NIPS 15*, 505–512.

[1] Recall that informativeness is different for each algorithm, while coherence is a property of the constraints and the distance metric; here, Euclidean distance was used.